CLAIMS

What is claimed is:

1. A system that facilitates incremental web crawls comprising:

an indexer that places items with similar properties into respective chunks; and,

a chunk map that stores at least some of the properties associated with the respective chunk, the chunk map employed to facilitate an incremental web re-crawl.

2. The system of claim 1, the items comprising information associated with a Uniform Resource Locator.

3. The system of claim 1, the items comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.

4. The system of claim 1, the items receives from a crawler, the crawler responsible for a specific set of Uniform Resource Locators.

5. The system of claim 1, further comprising a master control process that can modify the chunk map to facilitate load balancing amongst a plurality of crawlers.

6. The system of claim 1, further comprising a master control process that serves as an interface between a crawler and a re-crawl controller.

7. The system of claim 6, wherein the master control process maintains a known chunks table that stores information for components of a system.

8. The system of claim 6, wherein the master control process exposes an interface for communication with a component of the system.

9. The system of claim 8, wherein the interface returns a list of chunks the component should have and where to get the chunks.

10. The system of claim 8, wherein the interface returns a list of the chunks that should be actively served by the component.

11. The system of claim 8, wherein the interface returns a range of chunk identifiers to use in building a new chunk by the component.

12. The system of claim 8, wherein the interface causes an old chunk to be retired by the system.

13. The system of claim 6, wherein the master control process facilitates movement of chunks from one component to another component.

14. The system of claim 13, wherein movement of chunks is based, at least in part, upon at least one of rebalancing index servers after one goes down, re-crawling pages previously crawled, and, restoring a state of a crawler after it has crashed.

15. The system of claim 1, further comprising a re-crawl component that employs the chunk map to determine which chunks, if any, to re-crawl at a particular time.

16. The system of claim 15, the determination of which chunks to re-crawl, if any, being further based, at least in part, upon at least one of average time between change and average importance of documents comprising a particular chunk.

17. The system of claim 1, further comprising an index chunk that stores information associated with an index of at least some of the items.

18. The system of claim 1, further comprising a rank chunk that stores a static rank associated with an index chunk.

19. A method of performing document re-crawl comprising:

parsing a first chunk for uniform resource locators;

re-crawling the uniform resource locators; and,

forming a second chunk based, at least in part, upon the re-crawled uniform resource locators.

20. The method of claim 19 comprising at least one of the following acts:

determining whether any chunks are to be retired;

moving the first chunk; and,

destroying the first chunk.

21. One or more computer readable media having stored thereon computer executable instructions for carrying out the method of claim 19.

22. A method of performing document re-crawl comprising:

accessing a chunk map containing properties associated with respective chunks of data as a result of one or more web crawls; and,

periodically determining, based on the properties in the chunk map, whether to re-crawl one or more of the chunks of data.

23. The method of claim 22, the period determination being based, at least in part, upon, at least one of average time between change and average importance of documents comprising a particular chunk.

24. A data packet transmitted between two or more computer components that facilitates document re-crawl, the data packet comprising:

a chunk header that includes metadata associated with the data packet;

an offset section that provides offset information associated with document files; and,

the document files that include content found on the Internet.

25. The data packet of claim 24, at least one of the document files comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.

26. A system that facilitates increment web crawls comprising:

means for placing items with similar properties into respective chunks; and,

means for storing at least some of the properties associated with the respective chunk.

27. The system of claim 26, the items comprising information associated with a Uniform Resource Locator.

28. The system of claim 26, the items comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.